Indian Health Service (IHS)

Pilot Data Warehouse (PDW)
Task: FF141JS118T3

IHS DW-1 User Requirements/ Conceptual Design Presentation

Prepared For IHS By

September 17, 2001

Version 1.0

Document Title:

IHS DW-1 User Requirements/ Conceptual Design Presentation

Pilot Data Warehouse (PDW) Project -- Task: FF141JS118T3

Document Owner: Stan Griffith, M.D., IHS, Medical Informaticist (stanley.griffith@mail.ihs.gov)

Document Author: Betsy Miller, IBM Global Services, Business Intelligence Consulting & Services (betsym@us.ibm.com)

Document Create Date: 9/17/2001

Record of Revisions

Ver.	Modified by	Revision Date	Revision made
1.0	Betsy Miller Stan Griffith	9/19/2001	Final edits to User Requirements section.

Distribution

This document has been distributed to:

Version	Date	Name
1.0	9/17/2001	Stan Griffith, Stephanie Klepacki

TABLE OF CONTENTS

INTRODUCTION

Purpose of this Document

This document outlines the business and technical requirements and the high-level conceptual design for the IHS data warehouse project phase referred to as "DW-1". DW-1 is defined broadly as the first iteration (or release) of a production data warehouse system at IHS which will ultimately become an enterprise-wide data warehouse (EDW) supporting all of IHS' business intelligence strategies. Subject area content for the DW-1 release includes patient registration and healthcare encounter data across the IHS healthcare delivery system as far back as FY1997. The DW-1 requirements outlined in this document serve as an extension of requirements previously articulated during the pilot data warehouse (PDW) phase which is currently in development at IHS.

Requirements specifications serve as input to planning for design and development activities of each release of the data warehouse solution. Significant design elements for each phase are prioritized and driven by the objectives and scope outlined in the requirements statement. Thus, the purpose of this document is to validate the baseline requirements for the IHS data warehouse, and apply those requirements to the conceptual design specifications for the first iteration.

The document serves as a communication tool for planning the IHS data warehouse, providing information about the intended solution scope and direction, and how the data warehouse conceptual design aligns with stakeholder requirements. It can be used to solicit feedback and buy-in from stakeholders before proceeding with the production data warehouse design and development iterations.

Background: Pilot Data Warehouse (PDW) and DW-1 Project Phases

In the spring of 2001, IHS initiated a Business Intelligence project to design and build a data warehouse solution that will provide information, reporting and analysis resources to a range of agency stakeholders. The enterprise data warehouse (EDW) solution will include patient registration, healthcare encounters (outpatient visits, inpatient admissions, dental services, mental health, etc.), patient medical records, health risk factors, employees, contract healthcare service providers, third party eligibility coverage, financial information, and facility and equipment asset management.

The pilot phase of the data warehouse project (PDW) allows for the demonstration and testing of key technologies and processes related to data warehousing. These include data warehouse architecture, data modeling, extract, transform and load (ETL) processes, error checking and data quality improvement capabilities. In the pilot phase, a limited subset of patient and encounter data sources are acquired, loaded, and applied to a core set of population and workload reports, to serve as a test platform for the proof-of-concept demonstration. The pilot affords IHS an opportunity to test out the technological capabilities in a controlled environment, and refine the relevant design

elements before proceeding with iterative development phases for the full-production data warehouse solution.

The next phase of the data warehouse project (DW-1) calls for the extension of concepts developed in the pilot, using the same data subject areas, patient registration and encounters. The DW-1 project phase might include:

- a more inclusive set of data elements from sources such as patient registration, medical, dental and mental health encounters and admissions, medical record data, health risk factors, laboratory and pharmacy data
- additional data sources and source file formats (nation-wide coverage; standard and non-standard formats)
- a broader range of applications for using IHS data from the data warehouse (e.g. data marts)
- a larger number of users and broader range of user skill levels

For purposes of this requirements statement, it is assumed that DW-1 will be designed and developed as an enhancement to, rather than a replacement for, the PDW. Many of the requirements outlined herein are extensions of requirements already detailed in the pilot project phase. Until the PDW is developed, tested, evaluated and accepted as a pilot, (occurring concurrently with the development of this document), and then assessed against the stated DW-1 requirements, the nature and extent of re-engineering required for DW-1 cannot be determined.

Requirements Scope

Business and technical requirements drive the nature and content of each phase of the IHS data warehouse development project. Thus, the DW-1 project scope may be defined by a subset of the requirements contained within this document, those that are agreed upon and accepted in terms of priority, feasibility and timeline. These requirements may call for activities such as:

- Enhancement to the pilot data warehouse (PDW) architecture to improve performance or functionality and support database expansion
- Adaptation or reengineering of the PDW logical or physical data models to accommodate additional data element requirements or different entity relationships
- Enhancements to data cleansing and transformation routines based on research and findings about data quality and consistency
- Implementation of additional applications to support and enhance end-user accessibility and productivity
- Addition of dependent data marts or views to satisfy subject area-specific reporting and analysis requirements
- Introduction of the data warehouse to a wider user audience, including provisions for user training and user support
- Extension of metadata components to support expanded data subject matter, user audience, and applications

Specifically not in scope for the DW-1 phase of the IHS data warehouse are applications and data that might be required for healthcare encounter billing functionality, and reporting of other IHS subject areas such as employees, facilities, supplies and equipment asset management.

REQUIREMENTS DISCOVERY

Discovery Approach

The IBM Business Intelligence methodology has been adopted throughout the phased IHS data warehouse project, beginning with requirements gathering in the business discovery and infrastructure planning phases, followed by iterative phases encompassing solution definition, development, and deployment. This iterative process enables the introduction of an early release data warehouse that delivers value to IHS stakeholders for their immediate operational needs, while building incrementally toward a longer-range vision for the EDW solution.

As the PDW implementation proceeds, providing a proof-of-concept for key data warehouse technologies, the DW-1 business discovery and requirements definition activities proceed concurrently. This reduces the time required to fully implement the first subject area of the data warehouse solution, DW-1. However, it is important ot recognize that because DW-1 findings are in part dependent on findings from the PDW phase, the requirements specifications for DW-1 cannot be fully validated and finalized by means of this document.

User requirements for DW-1 are gathered through a combination of business discovery interviews and informal discussions with IHS stakeholders, including users, developers, statisticians, researchers, management analysts and administrators. Additionally, DW-1 requirements are gathered in work sessions throughout the planning, design, development, and evaluation stages of the PDW project. Finally, research is conducted from existing documentation related to IHS business objectives and strategies, as well as current challenges associated with information resources, data availability and data quality. Findings from these activities are organized into a Requirements Document, which provides input to the Conceptual Design for the DW-1 data warehouse iteration.

Users and Stakeholders

IHS information stored in the data warehouse may be used for a variety of business purposes by both internal and external stakeholders. Stakeholders include not only the direct users of IHS data warehouse applications, but also those who use the information sourced from the data warehouse for research and business functions. Additionally, the owners and suppliers of data for the data warehouse, and IT system developers and data administrators are important stakeholders in the formulation of requirements for the data warehouse solution. Interdependencies exist between the requirements for the IHS data warehouse solution and many other IHS data management initiatives currently underway or planned for the future.

Organizations and individuals that represent stakeholders, including direct users of the DW-1 data warehouse solution, may include:

- IHS Area Offices, Area Statistical Officers
- IHS/HQ divisions, including Office of Program Statistics
- IHS/ITSC organizations and systems such as RPMS and NPIRS
- IHS cross-functional work groups and task forces, such as the MPI Workgroup and the Data Quality Action Team
- IHS Senior Management
- Medical researchers, demographers and statisticians (both internal and external)
- Epidemiologists and IHS-sponsored Epi-Centers
- IHS-sponsored clinical programs, disease management programs, wellness and prevention programs
- Specialty care services such as Public Health Nursing, Dental, Mental Health, Chemical Dependency
- Accreditation programs, including ORYX
- Performance reporting programs, including GPRA (Government Performance and Results Act)
- Quality of care and outcomes measurement programs
- Tribal organizations and leadership
- Health care facility operations managers and analysts
- Contract health service providers, third party insurers, and fiscal intermediaries
- IHS administrators, fiscal managers, resource planning and funding allocation decision-makers
- External contractors and vendors providing data management related services or products to IHS
- External agencies or individuals requiring IHS information (DOJ, CDC, public)

The Business Problem / IHS Opportunity Areas

IHS requires a centralized, standardized, accessible source of reliable data related to the national IHS healthcare delivery system. The data required includes information about:

- the provision of health care services (ambulatory visits, inpatient admissions, education, diagnostic tests and results, immunizations, medications, etc.)
- patient diagnoses and other medical record information (medical history, health risk factors, problem list, etc.)
- patient population information (registration, demographics, eligibility, encounter history)
- financial information related to the provision of health care services (amount charged, amount paid).

The National Patient Information Reporting System (NPIRS) has provided a centralized repository of IHS data from a variety of sources. It has been used historically as the primary resource for generation and verification of annual user population and workload counts for the IHS healthcare delivery system, which supports decisions regarding the allocation of funding across IHS. However, problems with the NPIRS system have come under recent scrutiny.

Data in the NPIRS database is not received consistently nor on a timely basis across the IHS healthcare delivery system. The NPIRS database tables are structured in a sub-optimal denormalized manner that is cumbersome to use for basic query and reporting functions. No reliable universal identifiers exist in the database for identifying unique patients or encounters. To date, the NPIRS database has not yet been integrated with data acquired from other external sources (such as HCFA, U.S. Census, and CDC).

Access to NPIRS data resources is limited. While some reports are produced from the data in NPIRS, the database is not accessible to a wide user community for authorized ad hoc queries or user-generated reports. Data that may be accessed through NPIRS is often extracted and stored in divergent databases outside of NPIRS, resulting in higher costs of maintenance and inconsistent use of the data.

While NPIRS staff, working in collaboration with IHS statistical officers and the IHS Data Quality Action Team, have recently made substantial and tangible improvements to NPIRS data quality, much more work remains. User understanding of the nature and characteristics of IHS source system data is limited. Data is accepted into current repositories with limited error checking or analysis of data timeliness and completeness. Limited feedback is provided to suppliers of data regarding the nature of data received and processed. There is no integrated system for collecting, storing and disseminating NPIRS metadata (information about the data in NPIRS). Stakeholders need to better understand the data available to them in order to make the best use of information for management decision making and research purposes.

Known data quality issues, many originating from inconsistencies in source systems, have been thoroughly researched and documented. Other data quality issues are suspected and are currently under investigation. As issues are identified and resolved, users of the data are often unaware of the historical timing and impact of improvements made to data quality. Reports generated from the NPIRS database yield inconsistent, often inaccurate and unpredictable results. Data is NPIRS data structures remains unverified for extended periods of time, and user cannot reconcile NPIRS reports with those from others sources. As a consequence, users have tended to lose confidence in the NPIRS database as a reliable source of IHS patient care information.

With the design, development, and deployment of the new data warehouse, concurrent with ongoing initiatives to identify, analyze and resolve source data quality issues, IHS has an opportunity to create a valuable, credible information resource for its stakeholders. Users and administrators will come to depend on meaningful information from this nationally centralized source, to support funding justification, as well as new uses for healthcare information for clinical research and operations analysis purposes.

DW-1 Objectives

- Provide the capability to produce reliable and timely reports and data sets to support IHS statutory, regulatory and administrative obligations, including user population counts, workload reporting, accreditation and GPRA performance measurement.
- Improve the quality, timeliness and frequency of verified and published IHS data related to patient registration and health care encounters.
- Provide user accessibility to a comprehensive IHS healthcare database that will enable analysis and reporting of:
  - clinical practice patterns and episodes of care (diagnosis, treatment, wellness, prevention and screening services)
  - measures of quality of care, clinical outcomes, disease management, prevention
  - population-based epidemiological studies of disease states, medical histories, health behaviors, risk factors, and clinical outcomes
  - specialized healthcare programs including diabetes, dental, public health nursing, alcohol & substance abuse
  - patient demographics and healthcare utilization patterns
  - availability and accessibility of provider network service delivery areas
  - health care claim payment information, health care costs, budget and resource planning
  - insurance eligibility status and history
- Provide relevant information (metadata) about IHS data stores to enable end users to effectively leverage data resources for management decision making and clinical research.
- Standardize export programs, source data files formats and code values wherever possible, to maximize data processing efficiency and data quality.
- Provide user accessibility through web-based technologies and end-user desktop applications.
- Protect community, tribal and facility confidentiality in accordance with negotiated agreements and understandings with I/T/U organizations.
- Protect individual privacy and security of health care data, in compliance with applicable privacy regulations and IHS policies.
- Provide feedback to suppliers of data regarding detection of critical and non-critical errors affecting data quality and completeness.
- Restore stakeholder confidence in IHS centralized data sources and increase user self-sufficiency in accessing and using IHS data.

Critical Success Factors

It is important to consider the key factors that are critical to the success of the DW-1 project, and the potential risks associated with the project undertaking. Actions may be taken throughout the project phases to mitigate any potential risks and ensure the best possible outcome for the project.

Factors that are critical to the project's success include:

- Strong project leadership and committed executive sponsorship – Executive leadership, from both business and technical perspectives, is important to communicate project objectives and strategies, and to put in place appropriate resources that are available and committed to the project.
- Business user representation in defining objectives and requirements for applications and data marts – The investment in the data warehouse environment is driven by the needs of the

business community. A team of end-users from IHS programs and areas should be identified to work with the data warehouse/data mart design team to ensure that information is delivered in a form that is useful from the end-user perspective.

- Technical and business data warehouse design and development team with appropriate skills – the successful introduction of a data warehouse solution at IHS requires the collaboration of business and technical team members with specialized skills in business intelligence technologies and a clear understanding of IHS business processes. This team will ensure that the DW-1 design leverages appropriate technologies to deliver a solution that is valuable to the business.
- Clearly defined and documented business definitions, rules, and processes – Many data quality issues derive from a lack of understanding of the business data and how it is processed. Stakeholder frustration with national data sources can be overcome if the information in the data warehouse is understood from a business perspective. As business policies and processes are documented, the technical design and documentation of the data warehouse can follow, to ensure that users have a clear understanding of the information in the data warehouse and how to interpret it.
- Sufficient training and user support services appropriate for various user skill levels – Because the new DW-1 data warehouse will be introduced to a wider user audience, it is important to consider the additional support services that new users will require to become productive with the new system.

REQUIREMENTS SPECIFICATIONS

Requirements specifications are diverse and can be described at varying levels of detail, depending on the maturity of the system planning cycle and the extent of stakeholder participation in requirements definition. In this document, requirements are organized into several topic areas covering a range of business requirements for the DW-1 phase of the data warehouse.

Some requirements topics are described broadly at this early stage in the DW-1 project. These requirements will become more specific as the conceptual design evolves into the detailed solution design and development phases, and as users become more familiar with the features of the new data warehouse design. End-user functional requirements, for example, are described in general terms in this document, and can be specified in further detail at a later stage. Primary focus in this document is given to the design elements of the data warehouse architectural foundation. Greater attention is paid to detailing requirement specifications that will drive design decisions related to data content, database structures, data movement, and addressing the data quality and performance issues associated with the current NPIRS environment. This includes requirements topics such as data sources, data element content and meaning, transformation rules and cleansing of data, data confidentiality and security, and performance of data warehouse processes and applications.

Functional Requirements

Functionality of the data warehouse solution from an end-user perspective includes features such as:

- User accessibility – users require an intuitive interface that requires little training, provides a consistent "look and feel", and enables user self-sufficiency
- Web enablement – delivery of reports and query results to users via standard web interfaces
- Analysis capabilities, particularly multidimensional on-line analytical processing (MOLAP) analysis of data related to:
  - error detection in data processing
  - data quality reporting
  - healthcare utilization rates
  - demographic patterns
  - disease incidence patterns
- Ad hoc reporting capabilities - user-friendly tools for report creation and adaptation of existing reports
- Standard reporting capabilities – user generated requests for standard reports with user-specified filter and sort parameters
- Presentation capabilities – ability to format report display features and produce tabular, graphic charts, and geographic representations of data
- Vendor product compatibility – ability to use commercial off the shelf products (such as SAS and Crystal Reports) to generate queries, reports, charts and extracts

Data Content and Structure Requirements

Source Data (Exports)

While the pilot data warehouse included only a subset of IHS data sources, the DW-1 implementation is intended to include all available data sources for patient registration and healthcare encounter data. This includes creating re-exports of data from IHS area offices that did not participate in the PDW phase, and updating the re-exports for those that did participate. (Phoenix, Nashville, and Albuquerque areas participated in the pilot.) Additionally, non-standard format re-exports received directly from non-RPMS tribal delivery systems will be introduced into DW-1. (Penobscot was the only tribal-direct re-export received and tested in the PDW phase.) *Non-standard data sources such as tribal-direct and non-RPMS system sources required for DW-1 are listed in Appendix A.*

Furthermore, data is required for DW-1 from additional system sources such as those supporting mental health and chemical dependency services. These sources were not included in the pilot data warehouse, and may require specialized data security procedures in DW-1 for protection of patient privacy. The source file formats, ETL programs, and target data models for incorporating new patient encounter data sources will be introduced during the DW-1 project phase. Future data warehouse project phases will include provisions for adjusting ETL programs and target data models as new data sources or re-export formats are introduced.

For each additional data source included in the DW-1 phase, it is assumed that, like the PDW, historical encounter data will date back at least three fiscal years based on dates of service. Re-exports of historical data may be required where existing archives do not go back three fiscal years, or additional data elements are required for DW-1 that were not previously exported. (See Data Elements section)

Data Elements and Data Models

Modifications and enhancements to the PDW logical and physical data models will be required to optimize performance of transformation programs and data mart applications. Results from evaluation of the pilot will determine the nature and extent of re-engineering required. All of the data elements in the PDW model, *(provided in Appendix B),* are included as similar requirements for the DW-1 phase. Many of these data elements were carried forward from those in the existing exports of statistical and registration records (used for the NPIRS and ORYX data repositories) and entity/attribute relationships for those data elements were modeled for evaluation in the PDW project phase.

During the PDW phase an expanded list of required data elements was created with each new field classified as either “critical” or “nice to have” by selected users. Most of these new data elements were included as add-ins for the special PDW re-export programs and were also modeled in the PDW data model. It is intended that these same data elements will be incorporated into standard data export programs at some point in the future for all area and tribal exports. Thus they are included as requirements for DW-1. The DW-1 data model will build upon and enhance the PDW model, with incorporation of these new data elements and re-engineering as needed to model new

entity relationships, definitions, keys, etc. *A sample of the data model used for the pilot data warehouse is provided in Appendix C.*

Since many of these new data elements have not been used in national reporting or analysis in the recent past, they will be tested and analyzed for validity in PDW, and based on those results, may require adaptation for DW-1. Adaptation may not necessarily change how or when they are exported, but rather may involve how the new data elements are transformed and mapped into target fields in the DW-1 data model. Some may be deleted entirely from DW-1 requirements (because research and analysis in PDW reveals that they are no longer required in the data warehouse).

Some placeholders exist in the pilot data warehouse model, where target fields were built in the PDW database design, but many were either not populated from available export source data files or will be populated only using limited data to test the design. The data elements were modeled in the PDW phase in order to lay the groundwork for future enhancements to the warehouse environment. These and other new fields may be more completely modeled, sourced and populated at some point in the future, potentially during DW-1 implementation, when re-exports are modified to capture and send values for new data element requirements. Data elements that may need to be added or more completely modeled in the DW-1 data models include:

- Medications
- Prescription fill/refill transactions
- Referrals and authorizations for treatment
- Skin tests
- Laboratory tests and results
- Clinical measurements, vital signs, etc.
- Patient medical history data
- Patient health risk factors
- Problem lists from the medical record
- Financial information (costs, fees, billed amounts, paid amounts)

A more comprehensive incorporation of these data sources into the data warehouse will provide users a more complete view of patient care services, financial information, medical history, and patient demographics for research purposes.

Data Migration and ETL Requirements

A certain degree of processing is required for migrating source data into the PDW, in order to test out and provide proof-of-concept for data acquisition, ETL routines and error checking technologies. Additional ETL program development will be required in the DW-1 project phase, to complete the requirements for data preparation, transformation, cleansing, derivation and unduplication of data targeted for the DW-1 data warehouse solution. Furthermore, as user insight about the data and how it should be handled grows during the PDW phase, additional requirements may emerge that are applicable to DW-1 phase data.

Data Preparation

Data formats received from a variety of sources are processed, modified, and placed in a staging area, in preparation for transfer and loading into the PDW. Source files that were submitted in hierarchical or non-homogeneous record formats are converted to homogeneous formats, field delimiters are inserted into files where needed, and all records are tagged with information about their sources. Additionally, character fields that represent dates with 2-digit years are converted to date fields in DB2, (using business rules to determine the century by comparing the patient's date of birth with the date of service). These and other possible data preparation routines will be required for all new data sources and source file formats introduced in the DW-1 phase.

As a general principle and policy, IHS will strive toward standardization of source file formats, and establish standards for frequency and currency of data exports. Thus, many of the programs and processes developed to prepare data for the PDW will be applicable to the new data warehouse environment. However, DW-1 ETL processes will need to continue to support some non-standard export formats that may continue to come in from certain sources. Many of the tribal-direct data that was not included in the PDW phase may require customized ETL programming to incorporate non-standard data formats into the DW-1 warehouse. Each source format will be evaluated individually to determine the nature and extent of customized programming required.

Data receipt and data preparation processes that occur in PDW are logged and date/time stamped, and events that trigger the acceptance or rejection of source data files are communicated back to the supplier of the data. This process of notifying senders of files received, processed, and accepted/rejected must be replicated in the DW-1 environment.

Data Transformation and Source-to-Target Mapping

Data elements acquired from source files for the PDW database were mapped to target data elements and tables, with a limited set of transformation rules applied as appropriate. Each source data element was defined with business terms in the metadata and given a standard name in the target database. Likewise in the DW-1 solution, every data element introduced to the data model requires source-to-target mapping specifications defining how and where to migrate the data elements into the warehouse. If transformations or derivations of data values are required in the front-end ETL process, they are defined in the DW-1 source-to-target mapping as well.

Examples of data element transformation logic that is required for front-end ETL processing for DW-1 include the following:

- Parsing of data contained within one data element source into multiple data elements in the target file. For example, the 'Provider Affiliation/Discipline Code' is divided into separate codes, 'Provider Affiliation Code', and 'Provider Discipline Code'. Likewise, dental source records contain the fields 'ADA Code String' and 'Units String', which must be parsed into individual 'ADA Code' and 'ADA Units Quantity' fields respectively.
- Conversion of coded data values where enforcement of standardized codes is desired in DW-1. For example, the data element 'Sex' may be received with values of 1 and 2, or M and F. The transformation to standard code values may be required in the front-end ETL process, eliminating the need for users to manually convert these values in their reporting applications.

- Resolution of inconsistencies in code set usage might also be programmed into the DW-1 ETL process, where appropriate. Data elements may exist in source files with similar business meanings but having different domain code values. These codes may have been used during different time periods, from different source systems, or at different levels of granularity. Examples include the usage of old and new code sets for 'Cause of Injury Code', 'Place of Injury Code', 'Immunization Code', and 'Patient Education Code'. Depending on the code set, and the business user's understanding of the meanings of the codes, it may be appropriate to define a crosswalk table to transform old codes into new, or to roll up detailed code sets into higher levels of summarization. As a general rule, however, DW-1 requirements will not include resolution of these inconsistencies in code set usage. Rather, the DW-1 database will be modeled so that separate data elements are maintained, with distinctive labels and business definitions, to capture new and old code set values where they exist.
- Insertion of decimal characters where needed in ICD9 diagnosis codes and procedure codes. Industry-standard code schemas, defined by version 9 of the International Classification of Diseases (ICD9), contain a decimal character in a specific position within the 3 to 6-character codes. Because source data files do not always include these characters, transformation rules will be applied in DW-1 ETL to enforce consistency in the usage of decimal points in ICD9 data elements. Data received that do not contain decimals will have the "dot" character inserted in the appropriate position in the field before loading into the DW-1 database.
- Population of null, blank, or invalid field values with default or derived values. Some data source files occasionally contain records with certain data elements having null values, blank spaces, or invalid code values. In some cases, these records will reject in the ETL process because the absence of key data element values is classified as a "critical error". (See Data Cleansing below.) In other cases, however, the data element values can be transformed by deriving or defaulting a value for the data element. These types of transformations may be required in the front-end data load process in DW-1 if the data element is commonly used by all or many applications. Or, depending on the use of the data element, the transformation may be deferred to one or more downstream reporting applications or data marts, where it is applicable for a more limited set of reporting and analysis purposes. Examples of data elements identified by the Data Quality Action Team (DQAT) having invalid or missing values include: 'Clinic Code', 'Visit Type', 'Service Category', 'Location of Encounter', 'Disposition Code', 'Community of Residence', 'Discharge Date', and 'Length of Stay'. Transformation rules and requirements for derivations for each DW-1 data element must be specified and programmed on an individual basis.

Data Cleansing

Cleansing of data in the DW-1 data warehouse environment involves the systematic checking of records against specified data quality criteria. The error checking instructions may result in:

- rejecting the record prior to loading into the data warehouse, and notifying the sender that the record was rejected
- editing the record to correct for errors before inserting it
- accepting a "dirty" record into the database as-is

Errors that are detected and accepted into the database, either edited or as-is, may trigger warning messages for reports to the sender indicating that error conditions were detected.

Errors may be classified as "critical" or "non-critical". Critical errors are those that have missing or invalid values in key index fields, which trigger the record to reject during the front-end ETL processing. For example, if a unique patient identifier is needed for the primary key in the patient registration table, and that ID field is missing, the record cannot be inserted. The incoming registration record will reject as a "critical error". Depending on the data model design, keys and indices in the DW-1 data warehouse, the specifications for critical errors may change from what was defined as "critical" in the PDW phase.

Non-critical errors are detected during the front-end ETL data processing described elsewhere in this document (Data Preparation, Data Transformation, Unduplication of Data). Additionally, non-critical error checks might occur as data is migrated into individual data mart environments, if the error condition itself is relevant only to the users of a single data mart. In either case, the detection of non-critical errors may be programmed to appear in reports back to the sender of the data, depending on the nature of the errors.

Examples of non-critical error-checking rules required for DW-1 include:

- Date of Service/Date of Birth missing or invalid
- Date of Service prior to patient Date of Birth
- Discharge Date prior to Admission Date
- Missing Discharge Date (DW-1 policy may allow deriving this field if Admission Date and Length of Stay are provided)
- Length of Stay not equal to Discharge Date minus Admission Date (DW-1 policy may allow for missing discharge dates, so this error check may not be applicable.)
- Date of Service prior to export cutoff date (DW-1 policy to designate cutoff dates for encounters)
- Date of Service/Date of Birth in the future
- Date of Service/Date of Birth in 2-digit year format (DW-1 policy may allow accepting and converting 2-digit-year date fields)
- Missing/invalid values for required fields (DW-1 policy to designate fields that are required, including conditional requirements e.g. Quantity required for non-null CPT codes)
- Invalid/Inactive code values (code is not in IHS Standard Code Book or industry standard code sets, or code is not active on date of service)
- Invalid patient identifiers (such as SSN all 9's, HRN less than 6 characters, Patient Name = Demo, Patient)
- Invalid Units/Quantity values (numeric value is zero, negative, or out of prescribed range)
- Invalid Charge/Fee amount (numeric value is zero, negative, or out of prescribed range)
- Encounter Procedure or Diagnosis Code not valid for patient Age or Sex (DW-1 policy to determine invalid combinations e.g. Hysterectomy for male patient)

Some error checking edits and data cleansing rules are required specifically for User Population and Workload reporting. If such edits are neither applicable nor desirable for other users of the DW-1 data, they might be more appropriately positioned in a transformation routine that migrates data

from the data warehouse to a subject area specific data mart. Or, the rules might be built into specialized reporting program logic (without editing the data itself). Examples include:

- Conversion of Alaska ASUFAC/Location of Encounter Area Codes from 30 to 35
- Conversion of old numeric Place of Injury codes to new alpha codes
- Conversion of Dental encounter clinic codes from 56 to 39
- Derivation of Service Category (when missing or invalid)
- Derivation of encounter Type (when missing or invalid)
- Derivation of Location of Encounter (when missing or invalid)

Above are a few examples of error-checking rules to be considered for the DW-1 data warehouse. Each requires more detailed specifications for invalid conditions, cleansing conversion rules, as well as whether the error condition should result in rejecting the record, accepting the record with errors corrected, or accepting the record as-is. Each detectable error condition must be defined and classified so that they can be logged and reported in error feedback reports to data owners and suppliers.

Data Derivation

Certain required data elements in DW-1 do not come directly from source data files, but rather are created within either the data warehouse or data mart environments to support and enhance data reporting and analysis functions. Specifications are required for how values will be derived for these new fields required in the data warehouse/data mart data models. Examples of these data elements include:

- Unique Identifiers (surrogate keys) – New fields are required in DW-1 when unique identifiers are not provided from source systems to serve as the primary key in database tables. This occurs for tables storing both patient registration and encounter data.
- Encounter Categorization Codes – New derived fields may be required in either the data warehouse or specific data marts to indicate:
  - whether encounter services were delivered by a direct IHS or tribal provider, or a contracted health service provider
  - whether encounter services were provided in Inpatient, Outpatient or Dental delivery settings (previously distinguished in NPIRS by separating data into tables)
- Diagnosis Sequence Number – The order in which diagnosis codes are sent in an incoming encounter record must be stored with each code as they are loaded into the DW-1 diagnosis table. This enables users to identify the primary diagnosis for each encounter record when there are multiple diagnoses in a single encounter.
- Workload Reporting Encounter Flags – Criteria for whether an encounter record is reportable may vary depending on the nature and purpose of IHS reporting functions. Specific criteria are currently defined for "official" IHS Workload Reporting, but those criteria are subject to change over time. Flags are set for each encounter based on business rules that are built into derivation logic for setting the flags. Once these flags are set, the encounter may be counted, or not counted, in various report applications. If flags are derived in a data mart environment, the flag settings may be passed back to the data warehouse database for storage and use by other data mart applications. Current criteria for workload reportability includes the following rules:
  - Encounter record must be a non-duplicate
  - Date of Service (or Admission Date) is within the date range covered by the workload report

- Location of Encounter must be a valid, active APC facility (outpatient encounters only)
- Registering facility (ASUFAC) must be a valid facility (outpatient encounters only)
- Clinic Code must be a valid workload reportable clinic (outpatient encounters only)
- Primary Provider Code must be a valid workload reportable provider (outpatient encounters only)
- Encounter (Visit) Type Code must be I, T, O, 6, P, U, C or '?'
- Service Category Code must be A, S, O, H or '?'
- Tribal Code must be considered Indian (for the 'Indian Only' version of the Workload Report)

♦ Duplicate Record Flags – Since duplicate records will be accepted into the DW-1 data warehouse, flags are set in each record to indicate which records are duplicates. Business rules are required to drive the derivation logic for setting duplicate record flags to 'Y' or 'N'. Under most reporting circumstances, records flagged as duplicates should be ignored, since a corresponding "non-duplicate" record exists within the database to provide the relevant reportable data. (See 'Unduplication of Data' section below)

♦ Date of Last Encounter – This derived field would provide a more efficient and user-friendly approach to counting patients by a variety of user-defined criteria. Current "official" criteria for User Population reporting calls for counting registered patients who have had a "workload reportable" encounter in the prior 3 fiscal years. The processing logic for the annual User Population reports involves searching through historical encounter data to select out 3 years of workload encounters in order to derive patient counting criteria, and hard-coding the fiscal year time frame into the User Population report logic. By instead accessing a derived Date of Last Encounter value in the patient registration record, users may take a short-cut approach to counting patients directly from the patient registration tables. If users want to count patients more frequently or for different purposes, using something other than the 3 fiscal year criteria, they can freely do so by keying their query off a date range based on this Date of Last Encounter field.

Unduplication of Data

Known circumstances exist in which similar or identical records are exported from source systems and loaded to the data warehouse environment multiple times. Evidence of possible duplication of a record may indicate one of several conditions. It is important to distinguish between different conditions because each duplicate situation may call for different specifications for managing, interpreting and storing the data in the warehouse:

Specifications for "unduplication" may vary for different purposes. In some cases, an unduplication algorithm may be required to bypass duplicates entirely, such as when exact duplicate records occur in the same source file. Alternatively, duplicates may be accepted into the data warehouse and flagged as duplicates according to a pre-defined algorithm that also determines which of the duplicate occurrences is the "non-duplicated" record. Certain unduplication routines may be applicable as data is extracted, transformed and loaded into the data warehouse from its source file. Whereas other situations may call for specialized unduplication logic further downstream, such as at a single data mart level, where it is applicable only for a specific set of reporting applications.

Circumstances of data duplication might include:

♦ **A duplicate which is a "modification" record**, representing a correction to previously sent information. A patient's date of birth or gender, for example might be corrected by exporting a

modification to a previously erroneous patient registration record. Because these duplicate records are considered corrections to previously erroneous records, the new data will overwrite existing data in the Patient Registration table in the DW-1 data warehouse. This will occur on a limited basis, only for those duplicates (or corrections) that update a static table such as Patient Registration.

- **A duplicate which is a more complete or updated record**, intended to replace a previously sent record which was incomplete. Examples might be the completion of the amount paid financial field in an encounter record, or the addition of a previously unreported diagnosis code associated with an encounter. If updated encounter records can be identified as such in data exports, they can be applied to existing encounter records in the data warehouse, without retaining multiple copies (snapshots) of the redundant encounter data. Business rules would be required at the data element level for which encounter fields are candidates for updating. Alternatively, if updates to encounters cannot be identified as such, then encounter record "snapshots" might need to be retained and flagged as "non-exact" duplicates the DW-1 database (see below).
- **A duplicate which represents a "change in state" for a dynamic data element**, such as a change in a patient's eligibility status or patient address. These situations may not be considered true duplicates, since historical values for dynamic data elements are retained in DW-1 by design. Historical records are retained for each change in state so that reports may reflect the attributes in effect at a specific point in time. In PDW, and likewise for DW-1, data elements in the Patient Demographic and Patient Eligibility dynamic tables are captured with historical context. Duplicates of this nature require special handling in the data warehouse ETL process to prevent double counting. Specifically, a 'Start Date' field is captured and used in the primary key for each occurrence of a change in state, which ensures that each historical "duplicate" is retained, but remains unique and identifiable. No flags are needed to mark these records as duplicates since the 'Start Date' fields makes each record unique and provide historical context for the record.
- **A duplicate which represents a referral for an encounter,** which often appears to duplicate the encounter itself. Documentation of patient referrals often comes through in data exports in the form of encounter records, even though the referred encounter may never actually take place. If the referral does occur, the actual encounter record itself ultimately enters the data warehouse with similar, sometimes even identical data. (Dates of service, providers, and procedure codes may vary between what was originally referred and what was actually performed in the encounter). Referral data may be an important and relevant requirement for certain analysis and research purposes in the data warehouse. As referral records are exported to the warehouse, they should be identified and flagged as such, so that they are not mistaken for actual unique encounter events, nor as duplicates for existing encounter records. If they cannot be identified as they are exported, they may not be recognizable as duplicates in the warehouse, depending on the exact data element criteria for duplication flags in DW-1.
- **A duplicate which represents multiple units (quantity) for a single procedure code in an encounter.** This situation might occur when a single encounter involves multiple quantities of a single procedure code, such as a dental procedure performed on multiple teeth. As the encounter record is processed through various export, transform and load processes, the multiple quantity for the procedure code sometimes gets converted into single units of the same procedure code, repeated multiple times. As a result, the processed records may appear to be duplicates, because they reflect identical information (i.e. same patient, date of service, location, provider, and procedure code). Where possible, these records should be exported with each procedure code having a corresponding quantity (units), so that it can be treated as a single

encounter record in DW-1. Alternatively, if the duplicate records are identifiable as representing multiple quantities of a single procedure code in a single encounter, they can be consolidated into a single record, rather than flagged as a series of duplicates.

- **A duplicate where one or more patient identifiers is not unique.** This situation occurs when multiple patients are registered with identical patient identifiers such as social security number, chart number, or name. Duplicates of this nature will not be recognizable in the data warehouse, since unique Patient Registration IDs are assigned using other identifying information such as the ASUFAC and source of the registration record export. Thus, for example, many patient registration records can exist in DW-1 with the same SSN, some of which might in fact be duplications of the same person, while other occurrences of the SSN may represent distinctly different individuals. However, the assignment of a Master Patient Identifier (MPI), using specialized algorithms and other identifiers to recognize unique patients, may resolve many of the duplicate situations of this type.
- **A duplicate which is the result of missing data fields**, such as a blank patient identifier in an encounter record. Multiple records that appear to be duplicates may in fact be intended as distinct and separate records, but may be unrecognizable because of null or default values in identifying fields. For example, unique patient identifiers may be missing from some encounter records, so that encounters on the same date of service at the same location with the same provider might be indistinguishable. These types of records may reject as "critical errors" in the DW-1 load processing if the blank/null fields are required as identifying keys or indexes in database tables.
- **An "exact" duplicate which is bypassed (or overwrites existing data) during the data load process** -- An "exact" duplicate occurs when the same record, as defined by the values of specified data elements in the source record, is exported multiple times. For example, an "exact" duplicate encounter might be defined as having the same combination of Source ID, Encounter (Visit) ID, Patient ID, Date of Service, Location of Encounter, and Provider Code as an existing record in the data warehouse. Thus, the "exact" duplication would be recognized and bypassed during the data load since the record already exists in the data warehouse. The inclusion of Encounter ID in the definition of "exact" allows for preserving separate encounter records in the event of a patient seen twice by the same provider on the same day (the two encounters would have different Encounter IDs). Duplicates with "exact" matches may be overwritten, rather than bypassed, if precedence rules are established to determine whether or not the incoming source record data should overwrite the existing data in the data warehouse. Ultimately, only one copy of the "exact" duplicate data is retained in the warehouse.
- **A "non-exact" duplicate which is retained and flagged as a duplicate in the warehouse** – Based on the definition of "exact" duplicate (as illustrated above), other duplicate encounter records may be exported which are similar, but not "exactly" identical, to an existing encounter. For example, if the definition above applies for "exact" duplicates, then a "non-exact" duplicate might occur that has the same Encounter ID, Patient ID and Date of Service, but perhaps a different Source ID, Location of Encounter or Provider Code. In this instance, the duplicate is not "exact" and would not be bypassed or overwritten. Rather, it would be retained in the data warehouse as an additional "snapshot" of the encounter. Based on precedence rules, one or more of the duplicates would be flagged as such, while one instance of the encounter would be flagged as "not a duplicate" in order to accurately count occurrence of encounter events.

Certain trade-offs result from decisions regarding unduplication of redundant data in the DW-1 data warehouse. As instances of duplicate "snapshot" records multiply in the database, so do the

challenges related to query and report development and performance, as well as the cost of data storage and maintenance. The more narrow the definition of an "exact" duplicate record, the fewer the number of "snapshot" records that will need to be stored for each encounter. However, the trade-off is that some meaningful data about the encounter could potentially be lost if duplicate records are bypassed or overwritten. Even if duplicate encounter snapshots are always retained to prevent loss of potentially meaningful data, the data would be rarely used if it is flagged as "duplicate" data in the warehouse. It is important to understand business needs and uses for retained data in order to establish a reasonable balance between retention of relevant data, and discarding redundant data that is less meaningful or rarely used by the business.

Intelligent Error Reporting

Detected error conditions, whether critical or non-critical, must be logged and coded, and made available for reporting. So-called "intelligent error checking" is reported back to the data supplier, so that adjustments or corrections to business processes in source systems can be made. Error reports include summarized information, such as a count of error frequencies by type over a specified time period, so that trends and patterns may be identified. Or, more detailed information may be reported, such as a list of specified errors by individual source file and row number, so that error records can be corrected and resubmitted. In either case, logging and reporting intelligent error checking events provides a valuable tool in assessing and improving the quality of the data contained within the DW-1 data warehouse.

Data quality improvements in source data systems is also an important objective of intelligent error checking feedback reports. Data quality begins at the data source, such as in RPMS systems, where data entry occurs and standard code book tables are used for coding data. Quality assurance methods can be built into IHS business processes and source system edits to prevent "dirty" data feeding into the DW-1 ETL processing. As a general principle, DW-1 ETL processes will not substantially edit or cleanse data received from IHS areas, but rather will give the sender feedback about the nature and quality of data received. IHS HQ and Area Offices would use the feedback reports to work toward a goal of correcting data quality issues at the source facilities wherever possible.

The IHS Data Quality Action Team (DQAT) has been working throughout the year to identify data quality issues, research their origins, and propose and implement solutions. The team will continue to strive to resolve data quality issues that derive from either source systems or ETL processes. The team will also continue to coordinate DW-1 planning with other IHS task force initiatives, such as the MPI workgroup, so that the impact of these initiatives on DW-1 design and data quality can be assessed. As knowledge about the origins of data quality issues matures, so will the data cleansing rules, as well as the design of error-checking feedback reports.

The value of intelligent error reporting can be improved by allowing users to analyze errors and navigate through error report data with an on-line reporting tool. An OLAP data analysis tool can provide area and facility managers, for example, with meaningful measures of error types and frequencies for their own area or facility, and give the user flexibility to explore the relationships between different types of errors within their facility or across facilities. By helping users analyze

and interpret data error patterns, OLAP-based error reporting can be instrumental in achieving data quality improvement.

Security Requirements

Establishing provisions for data security in the DW-1 data warehouse environment is important to protect both data integrity and patient privacy. Since the information in the warehouse is an important IHS asset, access to the data is both a privilege and a responsibility. IHS wants to manage the security of data resources sensibly, so that users can make the best possible use of available information while maintaining adequate controls and protections over that data.

User Permissions and Responsibilities

Classes of users of data warehouse resources will be identified according to their job functions and specifically how their job functions relate to usage of information stored in the data warehouse environment. Based on these classifications, user groups will be assigned certain data access privileges that will meet user requirements, but also comply with both IHS data access policies and legal regulations. Within classes of users, data access privileges are individualized for each user based on specific user characteristics (such as which IHS Area Office a statistician represents, or what disease an epidemiologist is studying).

User access to the data warehouse itself, or access to an individual data mart application, is not necessarily granted for all of the data, all of the time. Permissions are granted to individual users for access to different levels and types of data within the application's data repository. This might be controlled by specifying rows or columns in selected database tables, or by permitting views at different levels of data aggregation (such as patient counts by selected dimensions). Furthermore, certain users might be granted data access privileges with both read and write permissions (including inserts, updates, or deletes), while most users will be limited to read-only data access privileges.

Policy and procedures will be established to ensure that users who are granted access to data understand their authorities and responsibilities with respect to the data, as well as the permitted uses for information acquired from the data warehouse. Because penalties for misuse of health care information are high, these policies must be executed responsibly and supported by technological capabilities within the data warehouse environment.

Access Controls

Controlling access to specific types of data may be required within the data warehouse database itself, or it may be controlled through functionality built into end-user reporting or analysis applications. At the database level, direct users of DB2 tables are restricted by pre-defined database views, which may control access to certain tables, or certain rows or columns within a table. Tables may be constructed at different levels of granularity, so that users might be granted access to only the summarized data within a table view.

- Record (Row-Level) Control

Individual records, or rows in a table typically represent individual occurrences of detail items such as healthcare encounters or patient registration events. Rows in a table may also represent counts, or aggregations of detail record-level data by certain pre-defined dimensions (e.g. encounter counts by month, by facility, service unit, and area). Access to data through row level controls allows users to view all rows meeting certain criteria. For example, Area Statistical Officers will require access to all of the health care encounters that were delivered within all service units and facilities in their area, but will not require encounter data from any other IHS Areas. Thus, an Area Statistical Officer might be given access to a table view restricted to the rows that contain his/her area's data, including the aggregated facility, service unit, and area-level counts. Data that is aggregated at a level higher than the area would not be accessible to the Area Statistical Officer, nor would any of the detail or aggregated data from any other area.

- Data Element (Column-Level) Control

Specific data elements, or columns in a table which identify individual patients, such as name or address, need to be restricted to protect the privacy of patients. For example, perhaps a user is granted permission to access any information in an encounter table, and can count encounter occurrences by any dimension (attribute) associated with either the encounter or the patient. For most purposes, the user would not need access to the patient's name, address or social security number, since the dimensions of interest for aggregating encounter counts do not normally include these attributes. However, since these attributes are stored in the same Patient Demographic table as other dimensions that are of interest (community of residence, for example), user access must be granted to the table itself, but restricted to only certain columns within the table.

- End-User Tools and Application Controls

Regardless of the database table, row and column permissions that control user access to specific types and levels of data, additional controls are required within end-user applications or data marts. Queries and reports in a data mart might be designed to enable more extensive analysis or manipulation of data, but provide more restrictive displays of information than might otherwise be controlled by direct-access to the database.

For example, a data mart might be created to manage health care data related to a defined population of patients. In one report application, users might be permitted to view a detailed list of patient identifiers and some attributes of the patients, such as their names or addresses. In a separate report, the user might be given permission to view detailed information about patterns of diagnoses for the defined patient population. But the user might not be permitted to identify which patients in the data mart had which diagnostic conditions. The applications developed in the data mart environment must be designed to support the necessary controls over data analysis, manipulation, and presentation in this instance.

Similarly, a data mart might be required that would provide users with a multi-dimensional, aggregated view of data, but limited access to more granular data. The user can navigate freely through aggregated data to analyze, interpret, and display meaningful information. However, the data mart application might control the level to which the user can drill down into detail or roll up

aggregated data. This might allow, for example, users to compare encounter counts by area across all areas in a data mart, but not allow drill down into comparisons at lower levels of detail across areas. It might be designed to allow users to drill down to further detail within one single area (the user's area), but not compare at detail levels across areas outside their own area. Finally, it might be designed to curtail drill-down to a designated cell level within the multi-dimensional data structure, or it might allow the user to "drill through" the multi-dimensional cube all the way into the most granular levels in underlying relational database tables. The design of the application and the security controls built within the data mart may vary depending on the intended purpose and usage of each application.

HIPAA and Privacy Act Compliance

Recently promulgated regulations from the Health Insurance Portability and Accountability Act (HIPAA), as well as those already in effect under the Privacy Act, will impact security requirements in the IHS data warehouse. HIPAA regulations apply specifically to all organizations that maintain or transmit electronic patient-identifiable health information. HIPAA's regulations were designed to protect the security and confidentiality of patient health information by regulating both the standards for interchange of electronic data, and the rights of individual patients with respect to:

- the transmission of patient identifiable health data,
- the uses for that data, and
- the release of protected health information to third parties.

The HIPAA regulations impact how IHS data warehouse administrators establish and maintain the integrity of stored health data, the authorization controls for access to the data, and the physical security of the data stores.

Appropriate safeguards must be in place under both HIPAA and Privacy Act regulations to prevent unauthorized release of patient identifiable information to third parties. It may not be possible or practical from a technology perspective to absolutely prevent such releases from the data warehouse or data mart environment. However, it is important to develop policies, procedures, training programs and technological strategies to minimize the potential for unauthorized release of information, and to detect when such releases have occurred. Users granted access to data warehouse resources should be informed of these policies, the importance of compliance as well as the sanctions or penalties for non-compliance.

Patient privacy rights protected by HIPAA may also impact the circumstances under which patient identifiable data in the IHS data warehouse may be used without prior notice to the patient. Under some circumstances, patient consent must be obtained before data can be used or disclosed, and the patient may have the right to revoke consent at any time. Furthermore, patients may have rights to view or update individual health data that is electronically stored or transmitted into or out of the data warehouse. Finally, if patient identifiable information from the data warehouse is disclosed to third parties, patients may have the right to request an accounting of those disclosures. Applicability of these patient privacy regulations to data in the IHS data warehouse should be carefully evaluated. If deemed to be applicable in the data warehouse environment, appropriate policies and procedures should be implemented to respond to and comply with patient consent orders, requests to view or update warehouse data, and requests for accounting of disclosures.

Monitoring Data Access

The DW-1 data warehouse environment must be designed to enable monitoring of user access and usage patterns to identify potential breaches of data security. IHS patient confidentiality and privacy policies, and regulations such as the Privacy Act and HIPAA, mandate that event logs be maintained that track certain data access and security events. IHS needs to know who logged into and accessed what data in the warehouse, and when those events occurred. Additionally, IHS needs to be able to detect and act upon unauthorized access or other potential breaches to data security policies. This can be accomplished through the management of event logs, and the creation of triggers that send alerts to a data warehouse security administrator when events meeting certain criteria occur.

Physical Data Security

Physical protection of the DW-1 database, hardware and software systems is an important requirement to ensure that information resources are consistently available to authorized users during regular business hours. A system security plan that is specific to the DW-1 business intelligence system environment is needed, consistent with IHS/ITSC business practices and policies. This plan should include provisions for the physical security of the facilities, hardware, and networking communications equipment. System and data back-up procedures should be developed and implemented, as well as regular archiving procedures for data that is not maintained on-line in the data warehouse environment. Procedures for recovering from system outages or disasters are needed that will ensure that the data warehouse system and network communications can be restored as quickly as possible and within a reasonable timeframe.

Metadata/Documentation Requirements

Like the data warehouse itself, metadata is developed and released in iterative phases, generally synchronized with the implementation of the data warehouse phases, and driven by the prioritized needs of both the system administrators and end-users. Metadata matures as the data warehouse matures. When changes are made to user requirements, data sources, applications and business definitions, the data transformation and load routines are changed, as is the supporting metadata. Users can adapt more easily to transitions in the data warehouse environment if metadata versioning is leveraged to track and publish data warehouse changes as they are introduced.

Effective metadata management requires a plan for creating and publishing metadata as appropriate for each stage of data warehouse development. While some metadata components were created in the PDW phase, they will need revision, replacement or enhancement as the DW-1 data warehouse is introduced. New users in the DW-1 environment will require a minimal core set of metadata components to use as a reference and help them get started using the new data warehouse.

Business definitions for data entities, attributes and code values in DW-1 are needed as an important component of the metadata documentation. Current versions of the IHS Standard Code Books are

incorporated into metadata, if those codes are the basis for standard DW-1 domain values. Non-standard codes used in any DW-1 data elements should be documented in the metadata. New code sets introduced in DW-1, such as the set of error codes created for intelligent error reporting, should also be incorporated into metadata. As codes are added or changed, metadata versioning should reflect the changes as well as the historical context for any code values present in DW-1.

Transformation and cleansing rules used in DW-1 ETL processing are also required in the DW-1 metadata. If any data element values are edited according to ETL business rules, those rules must be documented in the metadata so that end-users can understand what was done to source data loaded into the warehouse environment. As intelligent error-checking requirements change and mature over time, so does the metadata. Each change to metadata requires documentation of the effective date of the change so that users can interpret the resulting impact in the DW-1 data warehouse data.

Data Marts and Reporting/Analysis Requirements

The primary focus for the DW-1 phase is on establishing the data warehouse architectural technologies, and stabilizing the data sourcing, ETL, and error reporting processes, covering nation-wide patient registration and encounter data content. Once the data warehouse technologies are introduced and data content complete, extended requirements for one or more dependent data marts may also be considered in the DW-1 phase.

A dependent data mart is a subject area focused application which is dependent on data sourced from the DW-1 database. Thus, a data mart benefits from the standardized processing, cleansing and integration of detailed data from disparate sources, but can be tailored to support the specific analytical requirements of a subset of business users. Data processing redundancy is minimized by developing common data processing routines in the data warehouse, while allowing for more specialized processing in each individual data mart environment. Users of the data mart can rely on the consistency and integrity of the data stored in the data warehouse, and can leverage the metadata information provided in the data warehouse environment.

Data marts may consist of aggregated or pre-calculated views of data sourced from the most granular level detailed data in the data warehouse. Specialized tools or applications might be associated with a specific data mart depending on the intended business uses. User access permissions might be controlled within a data mart environment, separately from those of the data warehouse itself.

At IHS, data marts are required for a range of statistical and clinical analysis functions. Each can be designed and developed from the detailed patient registration and encounter data that is captured, organized and integrated in the DW-1 data warehouse. Stakeholders such as area statistical officers and epidemiologists can make use of one or more data marts to support their business processes, subject to security controls. The specific requirements for the IHS data marts developed from the DW-1 data warehouse are dependent on the business needs of the stakeholders.

For some IHS stakeholders, data marts may be merely a source for on-demand pre-formatted standardized reports that display aggregated information from refreshed data sets. While for others, a data mart may provide a robust analytical environment in which patient or encounter data can be explored by a variety of dimensions, aggregated to different levels, filtered by specified criteria, and presented in a variety of different formats such as tables, graphs or maps. More advanced analytical capabilities such as data mining and optimization may also be delivered in a data mart environment.

Focused data marts that might be developed from the DW-1 iteration include:

- **Intelligent Error Reporting and Analysis** – Error data that is output from ETL processes might be migrated to a user-accessible database enabled by an analytical tool such as a multidimensional OLAP application. Standard feedback reports about data errors can be produced and distributed from this data mart. However, management from the area offices, HQ and ITSC can go deeper into the report data, using the data mart to explore, interpret and analyze error results to identify data quality issues, patterns and trends.
- **Statistical (User Population and Workload Reporting) Data Mart** – The annual verification process for user population and workload statistical data can be brought into a data mart environment where it can be performed on a more frequent basis. Unique derivations and edits to the data that are required to support the specialized reporting criteria might be done in this data mart environment without impacting data warehouse content. An OLAP application in this data mart would allow users to quickly generate verified workload and user population counts by a variety of selected dimensions, filters, aggregation levels and time periods.
- **Epidemiology Data Marts for Defined Populations** – An epidemiological program might call for the development of a data mart application that is tailored for a particular epidemiology study, disease state or patient population. Selected data for a defined patient population might be used to populate a data mart view. Applications that enable epidemiologists to explore and display information would be developed for analysis such as disease incidence rates, utilization rates, baseline measurements, longitudinal studies, demographic trends, statistical correlations, clinical practice patterns, episodes of treatment, outcomes, morbidity/mortality patterns, or geographic (spatial) patterns of disease. Users might also incorporate additional data acquired from external sources (e.g. CDC, US Census, HCFA, state vital statistics) into the epidemiology data mart environment. Customized reporting as well as user access controls can be developed to meet particular security and patient confidentiality requirements in epidemiological data marts.
- **Accreditation Data Mart** – Features and functionality of the current ORYX database application can be replicated in a dependent data mart environment, with specialized reports and analytical tools to support accreditation business functions. Users can leverage the data quality achievements and processing performance improvements by accessing DW-1 data warehouse data in the development of a data mart to support accreditation analysis and reporting.
- **GPRA Reporting Data Mart** – Many of the IHS clinical performance indicators associated with the Government Performance and Results Act (GPRA) may be measured nationally from the centralized DW-1 database. A GPRA data mart could be created that would produce each of the measures using a standardized methodology. Data would be refreshed on a regular basis so that performance improvement trends can be followed over timeframes appropriate for each measure. Historical data in DW-1 can be applied to current performance measures in order to establish baselines retroactively. The measures would be supported by underlying detail data that would enable more in-depth analysis of results at national, as well as area and local site levels. A Balanced Scorecard application can be employed in the data mart that also presents

GPRA results graphically relative to standards and targets (e.g. traffic-light red/yellow/green indicators), allowing users to drill into any measured result for additional detail.

- **Public Health Nursing Data Mart** – The existing set of reporting applications that are produced specifically for the Public Health Nursing (PHN) program might be transferred from NPIRS to the new DW-1 data warehouse environment. Because these reports provide specialized views of public health nursing encounters and are used specifically by the PHN program, they may be packaged into a DW-1 based PHN data mart. Additional specialized derivations, calculations, or aggregations of data might be required, or tools for presenting and distributing custom reports might be used by the PHN program in a data mart environment.
- **Dental Data Mart** – Special requirements exist for reporting and analysis of data related to dental encounters and patient registration. A data mart specifically designed to provide dental program management with relevant data, reports and query capabilities could be created. Details for the requirements design and functionality of this data mart will depend on further business discovery with dental program management.

CONCEPTUAL DESIGN TASK

The conceptual design will document an architecture overview, data flows, descriptions of required transformations, project plans, cost estimates, and resource requirements. The conceptual design provides a high-level view of the DW-1 solution and describes how it achieves the benefits identified in the User Requirements discovery process.

APPENDICES

Appendix A -- Non-Standard Data Sources for DW-1

Appendix B -- Data Elements for Pilot Data Warehouse

Appendix C -- Pilot Data Warehouse Data Model